A Bayesian Analysis of Complex Dynamic Computer Models

Stefano Conti[*] Clive W. Anderson Anthony O'Hagan
Marc C. Kennedy

Department of Probability and Statistics
Centre for Terrestrial Carbon Dynamics
University of Sheffield

Abstract

Introduction

In many fields of science and engineering, the implementation of sophisticated mathematical models within large computer codes for simulating and predicting complex real-world phenomena is nowadays of widespread practice. Unfortunately the exploratory ability of such computer simulators is often hindered by considerable model preparation and computational requirements. Even in cases where the computational burden is not remarkably cumbersome, various uncertainty factors can still significantly compromise the performance of a computer model. Among the main sources of uncertainty affecting the processes of model building and validation are: parameter uncertainty, originating from unknown quantities tuning the code; model inadequacy, due to necessarily imperfect fit to the observed data; residual variability, related either to intrinsic randomness or unrecognised features of the real-world phenomenon; parametric variability, arising from quantities conveniently left unspecified; observation error, caused by inaccuracies at the data recording stage; and code uncertainty, related to the complex nature of the model. See Kennedy and O'Hagan (2001) for a thorough discussion of these points.

Several methodologies, useful to assess the reliability and effectiveness of a given computer model, are made available from the statistical literature:

[*]Department of Probability and Statistics, The Hicks Building, University of Sheffield, Sheffield S3 7RH, UK. Tel: +44(0)114 222 3824; fax +44(0)114 222 3759; e-mail: s.conti@sheffield.ac.uk

an exhaustive reference is provided by Saltelli et al. (2000). Unfortunately standard uncertainty/sensitivity analysis tools often require a large number of code runs, hence proving unsuitable for the validation of computationally expensive models. To circumvent this problem, an approach based on preliminary emulation of the code's outcome (*meta-modelling*) was suggested (Sacks et al., 1989) and later on largely pursued. This procedure would typically be followed by application of the aforementioned techniques to the emulator, which in fact is treated like a cheaper alias of the original code. In this setting, interesting results were obtained via a Bayesian semi-parametric representation of outputs from deterministic codes, that is models returning the same output when run over the same set of inputs. In previous works in the field (refer e.g. to O'Hagan et al., 1999) a Gaussian process prior distribution for the code's output was shown to be a convenient, flexible and often reasonable tool, especially for engaging the problems of model calibration and validation.

From static to dynamic models

Features in the structure of the code's input space may however complicate direct application of such methodology. This is usually the case with dynamic computer models, which are typically employed for examining time-evolving phenomena. Within such codes some of the inputs required during a stage of the simulation are actually outputs from previous stages.

In principle such codes can still be emulated in a single-run fashion: at the beginning of the time period of interest the model takes a specific set of inputs; at the end of it returns a collection of outputs. Under this perspective dynamic codes can be naturally accommodated within the statistical framework already existing for static models, no specific adaptation being required. Nonetheless non-negligible complications arise when applying this methodology: firstly, for increasing simulated time span the dimension of both the input and output spaces becomes less tractable, in that they have to encompass all time varying quantities; secondly, as the time period of interest changes the emulator needs to be rebuilt from scratch; in addition, model correction can be performed only at the end of the simulation.

An alternative approach to meta-modelling dynamic codes is instead based upon recognition of the stepwise nature of such models. After identifying a suitable basic time-step over which the simulator operates, a single-step emulator can be built; the latter is then run recursively in order to cover the simulated time period, in fact reproducing the intrinsic dynamic nature of the simulator. Stepwise emulators are usually characterised by more manageable input and output spaces than their single-run counterparts; moreover they enable interactive data assimilation, due to their improved flexibility.

A Bayesian viewpoint

In the present work a step-wise extension of the methods developed by O'Hagan et al. (1999) for static computer models is attempted. Basically the single-step emulator is assumed to be distributed *a priori* as a Gaussian process. The ensuing posterior distribution for the computer code outputs, given appropriately selected runs of the model, can be utilised to fully describe the original code after recursion is accomplished. Implementation of customary uncertainty/sensitivity analysis on the corresponding posterior moments is then expected to prove more efficient than direct MC routing of the simulator. Difficulties arising when running the emulator recursively comprise: assessing the meta-model's Normality, which in light of the stepwise nature of the emulation can at best hold approximately; ensuring no systematic drift occurs as the recursion spans the whole time period of interest; choosing a design set of input configurations adequately covering the model's input space; dealing with possibly high-dimensional hyperparameters. Although dimension-reducing strategies were devised in order to mitigate the emulator's computational load, it is still likely that a feasible Bayesian analysis of a dynamic computer model will impose some simplifications. Nonetheless it should be remarked that the outlined framework allows to address all the previously listed sources of uncertainty potentially affecting a computer model.

Application to an environmental model

A comprehensive test bed for the above illustrated methodology is offered in particular by the Sheffield Dynamic Global Vegetation Model (SDGVM), developed within the Centre for Terrestrial Carbon Dynamics for the purpose of examining the distribution of carbon stocks and flows in vegetation and soils under changing climatic and atmospheric conditions. Input parameters of SDGVM comprise broad soil, vegetation and climate descriptors; outputs of the model include various measures of a site's carbon budget and miscellaneous environmental quantities. Here extensive statistical analysis of SDGVM is required mainly because of: lack of knowledge regarding the true value of various tuning parameters; inaccuracies in the available data (either fixed and time-varying) needed both to initialise the simulator and to keep it "on track" with the real world; model's misspecifications due to incomplete scientific knowledge or understanding. Additional challenges presented by SDGVM comprise a high-dimensional input space and the existence of embedded sub-modules interacting at different time scales. When integrated with the developers' opinion on any relevant aspect of SDGVM, the Bayesian approach enables all the existing uncertainties to be effectively tackled, supporting modellers in both parameterising SDGVM and revealing incoherencies in the model's functioning.

References

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(3):425–464, 2001.

A. O'Hagan, M. C. Kennedy, and J. E. Oakley. Uncertainty analysis and other inference tools for complex computer codes. In *Bayesian statistics, 6 (Alcoceber, 1998)*, pages 503–524. Oxford Univ. Press, New York, 1999.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–435, 1989. With comments and a rejoinder by the authors.

A. Saltelli, K. Chan, and E. M. Scott, editors. *Sensitivity analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2000.